THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Word-aligned parallel text : a new resource for contrastive language studies

Citation for published version:
Volk, M, Göhring, A, Lehner, S, Rios, A, Sennrich, R & Uibo, H 2011, Word-aligned parallel text : a new resource for contrastive language studies. in Supporting Digital Humanities, Conference 2011. Copenhagen, Denmark, Supporting Digital Humanities, Conference 2011, Copenhagen, Denmark, 17/11/11.

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Supporting Digital Humanities, Conference 2011

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 05. Apr. 2019

Word-aligned Parallel Text – A New Resource for Contrastive Language Studies

Martin Volk, Anne Göhring, Stéphanie Lehner, Annette Rios, Rico Sennrich, Heli Uibo

University of Zurich, Institute of Computational Linguistics
Binzmühlestrasse 14, CH-8050 Zurich
Switzerland
volk@cl.uzh.ch

Abstract

This paper describes the opportunities that arise from automatic word alignment for bilingual concordances and contrastive language studies. We introduce our parallel corpus of Alpine texts in French and German and our web-based alignment search system. We explain how we have reduced the number of erroneous alignments in the output by distinguishing between dominant and miscellaneous translations. We are currently in the process of extending the system to a new language pair Spanish-Quechua. This poses special problems because of the scarcity of resources for Quechua but also because of the wide typological gap between the languages.

Introduction

Translated documents in multiple languages (here: parallel documents) are highly regarded as valuable resources for various tasks in natural language processing. The most prominent application is Statistical Machine Translation that needs large amounts of parallel text as input to statistical alignment and subsequent translation. But parallel documents are also useful for tasks as diverse as word sense disambiguation, terminology extraction and cross-language corpus linguistics.

Fortunately, increasing amounts of parallel texts have become available. Most notably are the large parallel corpora produced by multilingual political organisations like the European Union (Europarl, Acquis Communautaire) and the Canadian Parliament (Hansards). Other large collections of parallel texts consist of film subtitles or technical manuals (for example, opus.lingfil.uu.se).

The usefulness of these resources for contrastive language studies has increased tremendously with the possibility to automatically align the texts on the word level. This is routinely done as part of Statistical Machine Translation by methods based on cross-language co-occurrence in sentence-aligned texts. Gradually this word alignment information has been recognized as a new resource for lexicon lookup in context. The Linguee service (www.linguee.com) is the best-known example. It searches parallel texts in the web (offline), aligns them on the sentence and word level and provides usage examples for English, French, German, Portuguese and Spanish. Part of the word alignment and the sorting of the search results is based on bilingual dictionaries, but the remainder is based on automatic (i.e. statistical) alignment. The statistical alignment naturally contains alignment errors, but has the big advantage that it displays examples of previously undocumented translation options.

Bourdaillet et al (2011) have investigated the issues of using automatic word alignment for a commercial bilingual concordancer. They discuss various (statistical) methods for filtering bad alignments (e.g. a noun aligned with a determiner) and merging translation variants (e.g. singular and plural nouns or different verb forms). The paper includes a detailed evaluation using 2000 queries on the 8.3 million English-French sentence pairs from the Canadian Hansard.

In this paper we present our own Align+Search system. We start by describing the process of compiling a French-German parallel corpus. We then describe the implementation of the system. We conclude with an outlook on extensions for new language pairs and new functionalities.

Text+Berg Alignment Search

Inspired by Linguee we have built a search tool for our own word-aligned parallel corpus. We have digitized all yearbooks of the Swiss Alpine Club from 1864 until today (www.textberg.ch). For the first 92 years the books were published with mixed articles in German, French and Italian. Since 1957 the Club has published the yearbooks in parallel French and German versions. We have collected all books in multiple copies, and cut one copy of each open so that we were able to scan all books with automatic paper feed. Afterwards we used commercial OCR (optical character recognition) systems to convert the scan images to text. There our processing started with automatic correction of the most prominent OCR errors. For example, the umlaut ‘ü’ is often misrecognized as ‘ii’. Therefore we check for every occurrence of ‘ii’ whether it is a valid German word form. If not, we convert ‘ii’ into ‘ü’ and check again (see Volk et al. 2011).

Afterwards we structure the text by identifying article boundaries based on manually corrected tables of contents. We then split the text into sentences and determine the language of each sentence automatically. This procedure allows, amongst others, the recognition of German quotations in French articles and vice versa.

Subsequently the corpus is part-of-speech tagged and lemmatized with the TreeTagger in its standard configurations for English, German and Italian. For French we created our own parameter file for the TreeTagger based on the Paris 7 Le Monde Treebank. German word forms that are unknown to the TreeTagger (and thus do not receive a lemma) are analyzed by Gertwol which results in a high lemma coverage for

Figure 1: Screenshot of the Align+Search system

German. In contrast lemma coverage for French is low which calls for a special French lemmatizer in the next version.

For the parallel books we search for cross-language article correspondences. Fortunately, the articles are nearly in the same sequence in the parallel books. We compare the author names and the article lengths in order to determine article alignment. In this way we have aligned 2900 French-German article pairs with a total of 4.9 million tokens in French and 4.3 million tokens in German.

In contrast to article alignment, automatic alignment on the sentence level turned out to be tricky. Length-based methods, like, for example, the Gale & Church algorithm, work badly since our texts contain more noise than other parallel texts, due to missed paragraph boundaries, misrecognized captions or other OCR artifacts. Therefore we have developed an MT-based sentence alignment algorithm called Bleualign. It relies on the automatic translation of the German text into French. The French translation is then compared with the French text. If the Bleu score between a sentence pair from automatic translation and manual translation is high, the source sentence coupled with the target sentence are regarded as a good alignment candidate. The test includes the assumption that the sentences in both German and French are in the same order. The Bleualign method has resulted in much better alignments and better SMT results (see Sennrich and Volk 2010). With this method we were able to align about 90% percent of the sentences in the parallel part of our corpus.

The aligned sentences are taken as input for statistical word alignment (Tiedemann 2011), which in turn forms the basis for the alignment search. Our search system allows searches for word forms or for lemmas. When searching for lemmas the query for German *Haus* will also yield the genitive form *Hauses* as well as the plural forms *Häuser* and *Häusern* and their translation correspondences.

In contrast to Linguee we have built a domain-specific alignment search (see kitt.cl.uzh.ch/kitt/alignsearch/). All our documents are from the domain of alpine texts. This enables the user to search for domain-specific terms such as German *Steigeisen* (EN: crampon) or French *sommet* (EN: summit).

Align+Search offers a display sorted after hits in the target language, with descending frequency. As an example, figure 1 shows a screenshot of the hits for the German query word *Kaufmann* which – in the translation examples – is either part of a mountain name, a family name, or translated as *commerçant* (EN: merchant) or *négotiant* (EN: trader). A more frequent query term as e.g. German *Sturm* (EN: storm) finds 195 hits for *tempête*, 14 hits for *vent*, 12 hits for *tourmente*, and 6 hits for *orage*. Such queries give a concise overview of the translation options and their respective prominence in the textual domain.

Implementation of Align+Search

The basic functionality of Align+Search is that it takes a word as input and displays all the sentences from the corpus containing the search word, together with their aligned translations where both the query word and its translation are highlighted. The user may choose between querying a word or a lemma in either language, French or

German. There are two options for sorting the output. The default is sorting chronologically by article. The alternative option is to sort the hits by translation variants and their frequency (as in figure 1).

The Align+Search system is implemented in MySQL and PHP. We chose to store the corpus in a MySQL database to make use of its structure and the quick searchability of an indexed relational database. Moreover, MySQL has become a de facto standard DBMS for building database-driven web sites in the open-source world. PHP was chosen due to being a server-side scripting language that gives easy access to a database via SQL queries.

The French-German parallel corpus was provided in XML format, and the information about word alignments was imported from a simple text file. All the corpus data was retrieved from XML files into MySQL data tables using the LOAD XML command that makes it easy to take advantage of the original XML structure.

We created data tables for all words of the German and the French part of the corpus, respectively. Each record is a token (word or punctuation mark) from the original corpus with its PoS tag, lemma and location information. The word records come in the order in which they appeared in the text. We also created data tables for all articles in the German and the French part of the corpus and for the alignment information.

The web interface

The web interface for Align+Search is written in HTML, CSS, PHP and Javascript. The program displays a search form together with pull-down menus for the search options. It establishes a connection with the MySQL database, sends SQL queries to the database, processes the query results and formats the query results as one or multiple tables.

After the user has submitted the query, all the sentences containing the query, together with their source information (yearbook, article number, title and author) and their aligned sentences (translations) are displayed. The search word and its aligned word or phrase are highlighted.

Note that we only retrieve those sentence pairs that are mentioned in the alignment file. That is, the number of matches output by Align+Search can be smaller than the number of occurrences of the search word in the corpus.

We also allow wildcard searches, where the wildcard symbol * (arbitrary number of arbitrary letters) can be placed anywhere in the word – in the beginning, in the middle or in the end.

Alignment selection algorithm

There are usually several rows in the automatically created alignment file that refer to the same occurrence of the search word, so the program has to select the best of those.

Let us look at the following examples from the alignment file. Sentence 10 in article 1 of the yearbook 1995 has resulted in the following automatic alignments

- *grösste = plus grand de*
- *grösste = plus grand*
- *grösste = plus*

Which is the preferred alignment? It is not always a decision between one-to-many alignments but often also a decision between many-to-many alignments. For example, the same sentence has also resulted in the following alignments

- *eine lange Tradition = longue tradition*
- *eine lange = longue*
- *Tradition = tradition*

Picking the best alignment for highlighting is a delicate process. We have developed the following priority rules for the alignment selection:

1. If there exists a 1:1 alignment where the Part-of-Speech tags in German and French are the same, then we highlight single words (e.g. *Sturm* and *orage*)
2. If there exists a 1:1 alignment where the PoS tags are not the same, then we highlight whole phrases (e.g. *grösste* and *plus grand de*)
3. If no 1:1 alignment is found, but there exists a 1:many alignment, then we highlight the shortest target phrase (e.g. *Hochgebirge* and *haute montagne*).
4. If no 1:1 alignment is found, but there exists a many:many alignment, then we highlight the shortest phrases on both sides (e.g. *eine lange* and *longue*).

This procedure works nicely but cannot solve the problem of miscellaneous word alignments which creep in the statistical alignment and represent a nuisance for the user. For example, when we search for the German verb *leuchten* (EN: *to shine*), we get many correct translation examples: eight translation examples of French *briller*, five examples of French *luire* and some others. But we also get an alignment with the French word *on* (EN: *they*) which is clearly incorrect.

> *Darüber **leuchten** die Firne und Eisbrüche von Damma und Rhonestock …*
>
> *Au-delà **on** voit briller les glaciers et les séracs du Dammastock et du Rhonestock …*

This is especially irritating since a correct translation (*briller*) is part of the target language sentence. Therefore we have developed a cleaning algorithm, which distinguishes dominant versus miscellaneous translations. The idea is to correct a spurious translation variant if a frequent translation variant is present in the same sentence. The problem – of course – is to find the correct

line between frequent (= dominant) and miscellaneous translations. We have decided to regard all translation variants which account for more than 1% of the hits as dominant translations with function words excluded. This has resulted in a big improvement in precision.

An example may demonstrate the quality of the output. When we search for German *Hütte* (EN: cabin; among the most frequent words in our corpus) we get correct alignments for French *cabane, refuge, chalet, hutte, abri, bâtiment* (listed here in decreasing hit frequency) which account for 1200 occurrences. In addition we get five incorrect alignments with one hit each (*couchettes, être, intérieur, par, table*). Most incorrect alignments are due to a missing correspondence in the target language. The statistical word aligner selects a target word even if there is no fitting one. For example, in the following sentence pair the French sentence does not contain a word that corresponds to the German word *Hütte*. Instead the word aligner picks the French word *chouchettes* as the most likely correspondence (which is the translation equivalent of German *Schlafkoje*).

> *Abwartend ziehen wir uns auf die obere Schlafkoje zurück... und wieder betreten drei Männer die* ***Hütte****, bald gefolgt von einer Partie junger Leute...*
>
> *Nous nous retirons sur les* ***couchettes*** *supérieures, quand trois autres hommes arrivent, bientôt suivis par une troupe de jeunes gens.*

Performance of the system

The Align+Search system runs on the University of Zurich web server. It is very quick to deliver its results. Usually, the concordance table is displayed within a second. Queries with many hits may take a few seconds. Even if the query leads to more than 50,000 hits (for a function word like German *und*, for example), the program finishes its work in about 1,5 minutes. We achieved this performance by indexing the tables, creating integer key columns instead of string, optimizing the SQL queries and the PHP code.

Alignment Search for Spanish-Quechua

As a next step we will build an alignment search system for the language pair Spanish-Quechua. Quechua is a family of indigenous languages in South America spoken by 10 million people mostly in Bolivia, Ecuador and Peru. Despite the large number of speakers, Quechua is losing ground against Spanish. Spanish is the majority language and dominates administration and education (Rios et al. 2009).

Since Quechua is mostly a spoken language, parallel texts in Spanish and Quechua are scarce, although a small number of official documents have been translated (e.g. the Peruvian constitution).

Still, for learning Quechua, for the scientific investigation of Quechua and for strengthening its reputation, it is of utmost importance to provide parallel resources which allow Quechua speakers and learners to check word usage, and to allow Quechua researchers to investigate the language in detail.

We have acquired a small number of bilingual books and brochures in Spanish and Cuzco Quechua, which will form the basis for our alignment search system. Since Quechua is a strongly agglutinative language, we need to split the words into so called 'inflectional groups', a procedure that has been described for the annotation of the Turkish METU-Sabanci treebank (Atalay 2003). An inflectional group consists of one or more morphemes that, in our case, give a good basis for the alignment to Spanish words.

We have already built a finite state analyzer for Cuzco Quechua morphology (Rios 2010). We also have bilingual word lists (more than 2000 base forms) so that the resources for a lexicon-based word alignment are given. In addition, we have commissioned the manual translation of German-Spanish parallel texts (~ 100,000 words) into Cuzco Quechua as part of our project to build a trilingual German-Spanish-Quechua parallel treebank. We have selected texts from the domains agriculture, nature, and education.

We have conducted several word alignment experiments with GIZA++ (which is part of the SMT toolkit Moses) on our Spanish-Quechua corpus. As for Quechua we tested alignment on words and on inflectional groups. The experiments confirmed that statistical word alignment from Spanish to whole Quechua words is very difficult: Then the error rate was around 80%. We achieved a clear improvement through the segmentation of Quechua words into inflectional groups. We hope to further improve the results as we increase the size of our parallel corpus and will be able to train GIZA++ on a larger amount of Spanish-Quechua parallel texts. We expect that automatic alignment will improve sufficiently to serve as a basis of the alignment search system.

As part of a field trip to the Quechua heartland in Cuzco, Peru, we searched for bilingual street signs and posters in both Spanish and Quechua. We include a few examples (figures 2 and 3). We intend to include these as visual underpinning of our alignment search system in order to ground the system in real world examples.

Conclusions

Because of automatic word alignment methods, it is now possible to provide large word-aligned parallel texts. These constitute a new way of accessing lexical data in context. Alignment search systems like Linguee or our Align+Search help not only in language learning but also in translation studies. They allow for the discovery of previously undocumented translation alternatives and word senses. When enriched with linguistic information such as named entity classes, syntactic trees or co-

reference tags, they open new possibilities for corpus linguistics.

The following extensions of the functionality of Align+Search are planned for the near future.

- We will enable reverse queries. For example, when querying for German *Seite* (EN: page, side) we get, among others, hits of the French word *versant*. A button-click will issue a query for *versant* that shows all its German translations in the corpus. In fact, we envision that this will help in semantic mirroring of the various word senses (Dyvik 2004).
- When querying for German nouns, we want to suggest the diminutive form (*Hütte : Hüttchen*) and the opposite gender form (*Wirt : Wirtin*) for future searches. Something like: "Are you also interested in X?" when such forms exist in our corpus.
- We will allow queries not only by a single word (or lemma), but also by multiple words, by word + PoS tag, and by sequences of words and PoS tags.
- We also think that it will be useful to indicate for each translation pair which is the source language und which is the translated language. This will allow to better appreciate the translation variants.
- We plan to include parallel texts from different genres and domains. This will allow contrastive views of translation options. For example, when the user queries for *Gipfel* (EN: *summit*), we can show which are the most important translation options in the Alpine domain and which are the most important in the domain of politics.
- Automatic alignment will always contain some erroneous alignments. We will allow the user to mark these errors and to suggest correct alignments. We will exploit the user input to improve the alignment quality of the system.

Acknowledgements

We would like to thank the many students who have helped in building the Text+Berg corpus. Special thanks go to Patricia Scheurer for compiling the table of source vs. target languages. We gratefully acknowledge that this research was funded by the Swiss National Science Foundation under grant 105215-126999 “Domain-specific Statistical Machine Translation”.

References

Nart B. Atalay, Kemal Oflazer, and Bilge Say (2003). The Annotation Process in the Turkish Treebank. In Proceedings of the 4th International Workshop on Linguistically Interpreteted Corpora (LINC).

Julien Bourdaillet, Stéphane Huet, Philippe Langlais and Guy Lapalme (2010). TRANSSEARCH: from a bilingual concordancer to a translation finder. In: Machine Translation. 24. pages 241-271.

Helge Dyvik (2004). Translations as Semantic Mirrors: From Parallel Corpus to WordNet. In Karin Aijmer and Bengt Altenberg (eds.): Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg. May 2002. Rodopi. pages 311-326.

Annette Rios (2010). Applying Finite-State Techniques to a Native American Language: Quechua. Licentiate thesis. University of Zurich, Institute of Computational Linguistics.

Annette Rios, Anne Göhring, and Martin Volk (2009). A Quechua-Spanish Parallel Treebank. In: Proceedings of 7th Conference on Treebanks and Linguistic Theories. Groningen.

Rico Sennrich and Martin Volk (2010). MT-based Sentence Alignment for OCR-generated Parallel Texts. In: Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010). Denver. 2010.

Jörg Tiedemann (2011). Bitext Alignment. Morgan & Claypool Publishers.

Martin Volk, Lenz Furrer and Rico Sennrich (2011). Strategies for Reducing and Correcting OCR Errors. In: C. Sporleder et al. (eds.): Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series. Springer-Verlag. Berlin. pages 3-22.

Figure 2: Pointer to the Main Auditorium at the Centro Bartolomé de las Casas, Cuzco. Hatun Rimanakuna Wasi = the big house for speeches

Figure 3: Sign at Cuzco bus terminal. Haykuna Punku = the gate for entrance